Sparse Representations

Sargur N. Srihari
srihari@buffalo.edu

Regularization Strategies

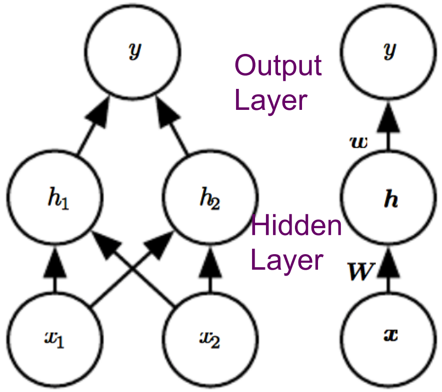
1. Parameter Norm Penalties
2. Norm Penalties as Constrained Optimization
3. Regularization and Under-constrained Problems
4. Data Set Augmentation
5. Noise Robustness
6. Semi-supervised learning
7. Multi-task learning
8. Early Stopping
9. Parameter tying and parameter sharing
10. Sparse representations
11. Bagging and other ensemble methods
12. Dropout
13. Adversarial training
14. Tangent methods

Direct and Indirect Penalties

- Direct Penalty
  - Weight decay penalizes parameters directly
  - L^1 penalization induces sparse parameterization
- Indirect Penalty
  - Another strategy is to place penalty on the activations of the units in the neural network
    - Encouraging their activations to be sparse
    - It imposes a complicated penalty on model parameters
  - Representational sparsity describes a representation where many of the elements of the representation are close to zero

Definition needs Matrix notation

- Given network drawn in two different styles
  - Matrix W describes mapping from $\boldsymbol{x}$ to $\boldsymbol{h}$
  - Vector $\boldsymbol{w}$ describes mapping from $\boldsymbol{h}$ to y
  - Intercept parameters b are omitted



- Layer 1 (hidden layer): $\boldsymbol{h}$ computed by function $f^{(1)}(\boldsymbol{x};\ W,\boldsymbol{c})=\boldsymbol{h}=g(W^T\boldsymbol{x}+\boldsymbol{c})$
  - $\boldsymbol{c}$ are bias variables
- Layer 2 (output layer) computes $\boxed{f^{(2)}(\boldsymbol{h};\boldsymbol{w},b)=\boldsymbol{h}^T\boldsymbol{w}+b}$
  - w are linear regression weights
  - Output is linear regression applied to $\boldsymbol{h}$ rather than to $\boldsymbol{x}$
- Complete model is

$$f(\boldsymbol{x};\ W,\boldsymbol{c},\boldsymbol{w},\boldsymbol{b})=f^{(2)}(f^{(1)}(\boldsymbol{x}))$$

Direct versus Representational Sparsity

– Parameter regularization

$$\underset{\boldsymbol{y} \in \mathbb{R}^m}{\begin{bmatrix} 18 \\ 5 \\ 15 \\ -9 \\ -3 \end{bmatrix}} = \underset{\boldsymbol{A} \in \mathbb{R}^{m \times n}}{\begin{bmatrix} 4 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & -1 & 0 & 3 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & -4 \\ 1 & 0 & 0 & 0 & -5 & 0 \end{bmatrix}} \underset{\boldsymbol{x} \in \mathbb{R}^n}{\begin{bmatrix} 2 \\ 3 \\ -2 \\ -5 \\ 1 \\ 4 \end{bmatrix}}$$

Weight matrix W is sparse

– Representational regularization

$$\underset{\boldsymbol{y} \in \mathbb{R}^m}{\begin{bmatrix} -14 \\ 1 \\ 19 \\ 2 \\ 23 \end{bmatrix}} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times n}}{\begin{bmatrix} 3 & -1 & 2 & -5 & 4 & 1 \\ 4 & 2 & -3 & -1 & 1 & 3 \\ -1 & 5 & 4 & 2 & -3 & -2 \\ 3 & 1 & 2 & -3 & 0 & -3 \\ -5 & 4 & -2 & 2 & -5 & -1 \end{bmatrix}} \underset{\boldsymbol{h} \in \mathbb{R}^n}{\begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \\ -3 \\ 0 \end{bmatrix}}$$

$\boldsymbol{h}$ vector is sparse

Representational Regularization

- Accomplished using same sort of mechanisms used in parameter regularization
- Norm penalty regularization of representation
  - Performed by adding to the loss function J, a norm penalty on the representation.
    - The regularized loss function is

$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{h})$$

where $\alpha \in [0,\infty)$

    - An L^1 penalty term induces sparsity: $\Omega(\boldsymbol{h}) = ||\boldsymbol{h}||_1 = \sum_i |h_i|$

Placing constraint on Activation Values

- Another approach to representational sparsity:
  - place a hard constraint on activation values
- Called Orthogonal matching pursuit (OMP)
  - Encode $\boldsymbol{x}$ with $\boldsymbol{h}$ that solves constrained optimization: $\boxed{\underset{\boldsymbol{h}, \|\boldsymbol{h}\|_0 < k}{\arg\min} \|\boldsymbol{x} - \boldsymbol{W}\boldsymbol{h}\|^2}$
    - where $\|\boldsymbol{h}\|_0$ is the number of zero entries of $\boldsymbol{h}$
    - Problem is solved efficiently when $\boldsymbol{W}$ is orthogonal
  - Often called OMP-k, where k is no. of zero entries
    - OMP-1 is very effective for deep architectures
- Essentially, any model with hidden units can be made sparse:
  - sparsity regularization is used in many contexts